Anja Zhang
INF1340 - Final Assignment Write-up

Final Assignment Write-up: Visualization & Analysis

Introduction

The Final assignment for INF1340 is based on utilizing data visualization principles on the previously cleaned dataset from the “Trends in International Migrant Stock: The 2015 Revision” published by the United Nation. The process will be using principles suggested by Edward Tufte and what he has developed as principles for the effective visualization of data. Some of these principles include

1. **Show comparisons:** Visualizations should allow the viewer to easily compare different data points or values.
2. **Show causality:** Visualizations should show the cause-and-effect relationships between different data points.
3. **Show multivariate data:** Visualizations should show multiple variables and their relationships, rather than just one variable at a time.
4. **Use appropriate encodings:** Visualizations should use the right visual encodings, such as position, length, and color, to represent different data values accurately and effectively.
5. **Use integrated captions:** Visualizations should include captions that provide information about the data and help the viewer interpret the visualization. Integration captions will be included with each visualization throughout the report.
6. **Show Context:** Describe or depict the before and after state. Show trend lines to hint at results in the future.

Edward Tufte's principles are widely considered to be fundamental guidelines for creating effective and informative data visualizations. By following these principles, the goal is to can create visualizations that are clear, concise, and easy to understand, and that effectively communicate the information and insights contained in your data.

Throughout this project, there is various types of visualization tools were deployed to perform visualization and analysis of the data provided. They include

1. **Bar charts:** These can be used to show the total population of different countries or regions. The bars can be sorted in ascending order to show which areas have the largest populations.
2. **Maps:** Maps can be used to show the distribution of a population across a geographic area. Different colors or symbols can be used to represent different population densities.
3. **Scatter plots:** These can be used to show the relationship between two variables, such as population and GDP. Each point on the plot represents a country, with the x-axis representing GDP and the y-axis representing the population.
4. **Boxplot chart:** These charts are the visual representation of the data set based on a five-number summary: the minimum, the first quartile (25th percentile), the median (second quartile), the third quartile (75th percentile), and the maximum. They typically look like a box.

5. **Violin plot chart:** These charts can depict the distribution of the dataset by using density curves. The width corresponds to the approximate frequency. They typically have a vertically oval shape and each end represents the maximum and minimum.
6. **Line charts:** These can be used to show how a population has changed over time. The x-axis can represent time, and the y-axis can represent the population.

Method & Result

1) The first set of data is showing information on international migrant stock at mid-year from 1990 to 2015. In this case, one insight in which we can derive from this dataset is what is the trend of international migrant stock throughout the years. Using Tufte's principles of showing comparisons, a bar graph chart was used to show the year-to-year comparison. Furthermore, this dataset also includes data on the sex of the migrants. Using Tufte's other principle of showing multivariate data, the sex of migrants was included in the y-axis to show if this trend is consistent throughout both sexes. Thirdly, based on Tufte's principle of showcasing causality, this graph showcases that as the year progresses, international migrants also have increased. The result of the bar graph shows there has been a consistent increase in international migrants from 1990 to 2015. This trend is also consistent with female and male migrants.

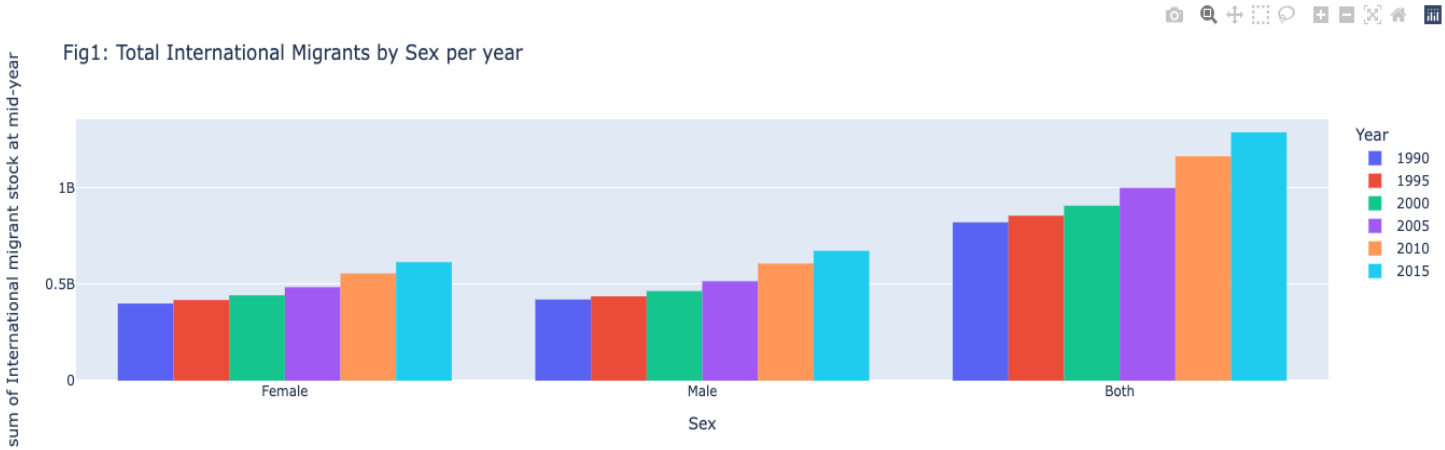


Figure 1: Bar chart displaying global total international migrants by sex. The x-axis shows the total international migrant stock and the y-axis shows the sex of the migrants. Each color represents a corresponding year in which the data was collected.

2) The second subset of data is showing information on the total population of over 200 countries across the world from 1990 to 2015. It would be challenging to display all countries in one chart therefore using Tufte's principle of show comparison, the best way to visualize such data is on a world map. It also uses appropriate encoding to showcase populations across the world. Using gradient colors so that the user can quickly decipher the range of the population. An animated feature is also included to display the time

lapse from 1990 to 2015. This animated feature can showcase any significant change in population with any color changes.

Fig2: Heatmap for Global Poulation

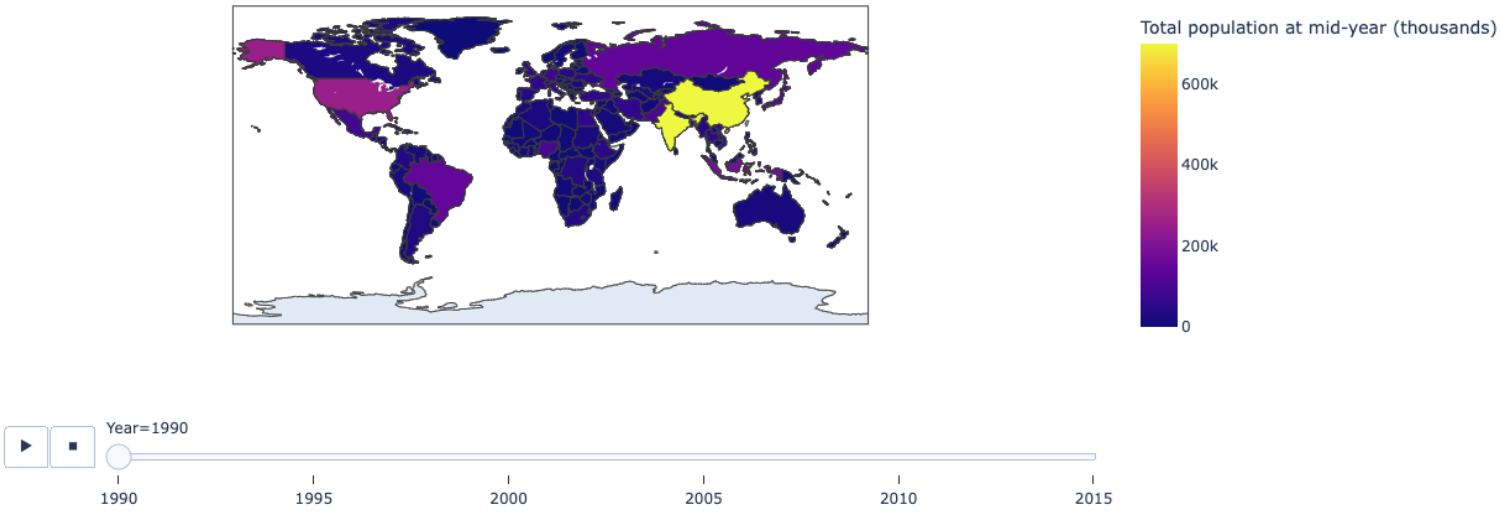


Figure 2: Heatmap which displays the total global population with the animated feature for timelapse through 1990 to 2016. The total population at mid-year (thousands) is color-coded using a gradient of colors. Deep purple means less populated and yellow means more populated.

3) The third subset of data is on information for international migrants stock as a percentage of the total population with the corresponding year and sex of the migrants. To visualize this data using Tufte's principle of showing comparison, causality, and multivariate data, a scatter plot graph was created alongside its trendline and marginal distribution plots to showcase its context. To draw a comparison between female and male international migrants throughout the years. The graph was created with the x-axis showing the year from 1990 to 2015 and the y-axis showing international migrant stock as a percentage of the total population. The color on the chart is also showing the data for female and male migrants. The results show that for the female migrant population by country, the maximum is 88.349%, the minimum is 0.033%, and the median is 3.399%, with the 25th percentile at 1.247% and the 75th percentile at 10.531%. Male migrant population by country produced a very similar result with a maximum of 90.697%, minimum of 0.032%, and median of 3.632%, with the 25th percentile at 1.355% and 75th percentile at 10.993%

Anja Zhang
INF1340 - Final Assignment Write-up

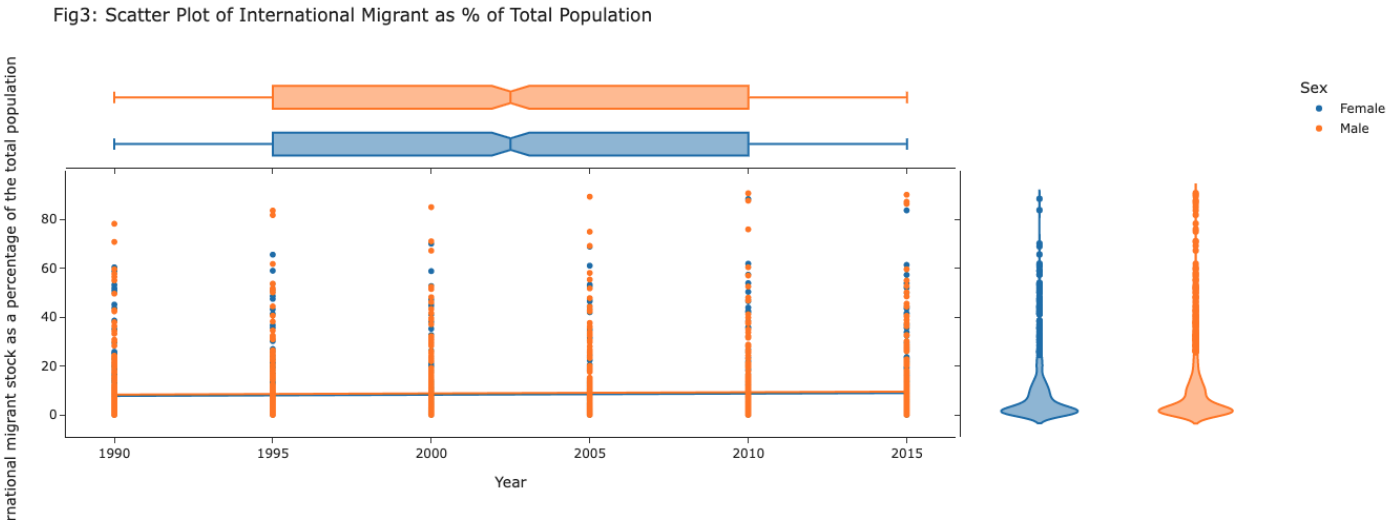
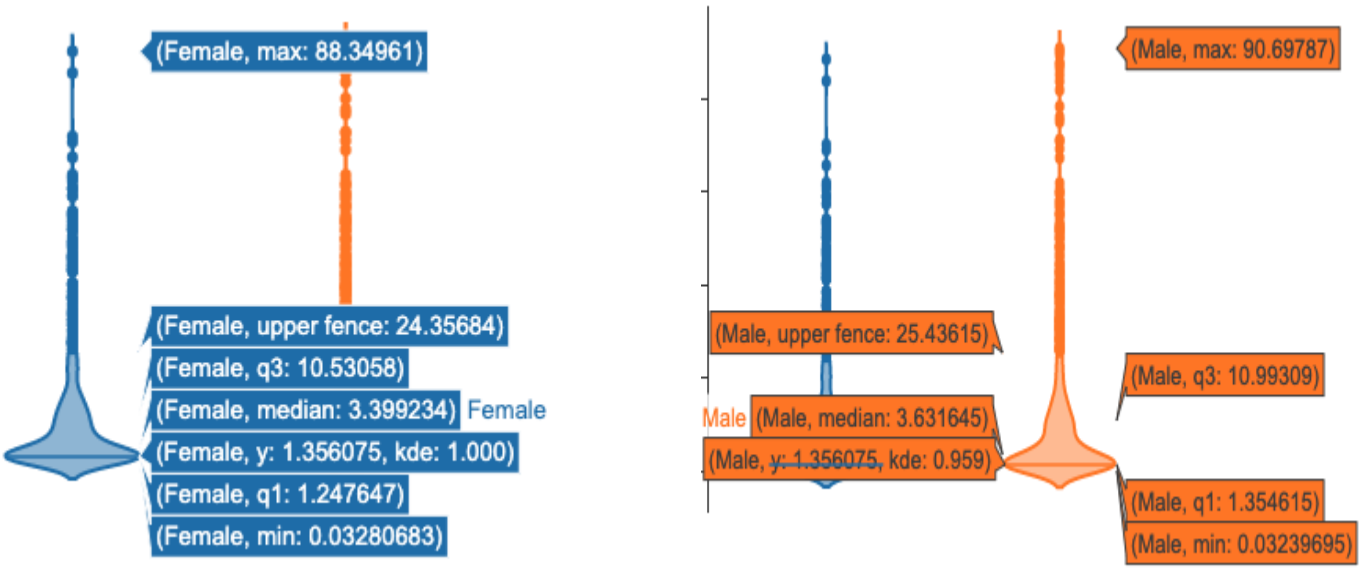


Figure 3: Scatter plot graph showing international migrants as a percentage of total population. The X-axis shows the year and the y-axis shows international migrants as a percentage of the total population. This graph also includes marginal distribution for female and male migrants.



Marginal Distribution in Figure 3: Showcasing the distribution for the female and males of international migrant stock as a percentage of the total population.

4) The fourth subset of data is on information for female migrants as a percentage of the international migrant stock from 1990 to 2015. Using Tufte's principle of showing comparison, causality, multivariate data, and context, the boxplot graph was created to showcase the year-to-year comparison for any significant fluctuation of female migrants. Using its hover feature, the user can find out at a glance what countries in which are the outliers in this context. This graph used years for its x-axis and female migrants as a percentage of international migrant stock as the y-axis. With this setup, it's very apparently the year-to-year change for female migrants. The results are displayed for the following data:

	Min	Q1	Medium	Q3	Max
1990	13.857	45.901	48.976	51.487	70.704
1995	13.858	46.463	48.965	51.913	68.549
2000	13.859	46.228	49.238	52.156	66.296
2005	13.628	45.961	49.123	52.241	65.384
2010	13.458	45.873	49.322	52.077	67.249
2015	13.326	45.821	49.438	52.184	68.964

Throughout the result, we can see that female migrant for all data points has been slowly increasing their percentage for international migrant stocks. We can infer that by following this trend we can see that there will be half migrant population to be female shortly.

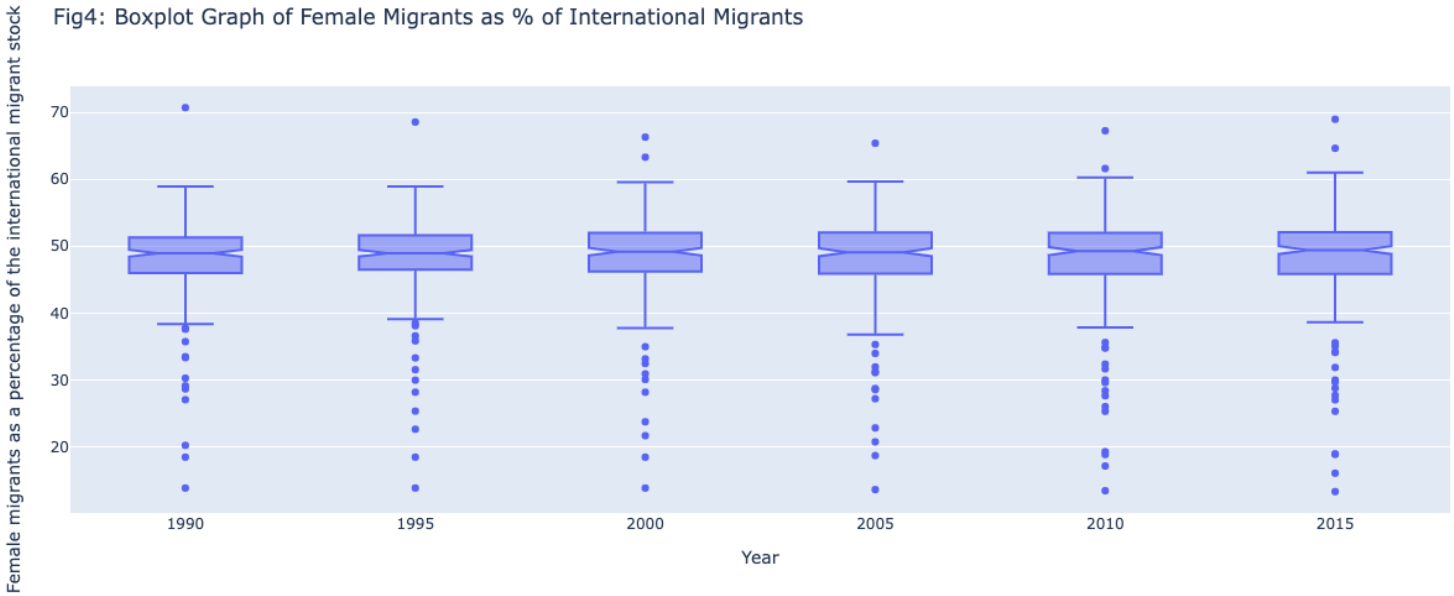
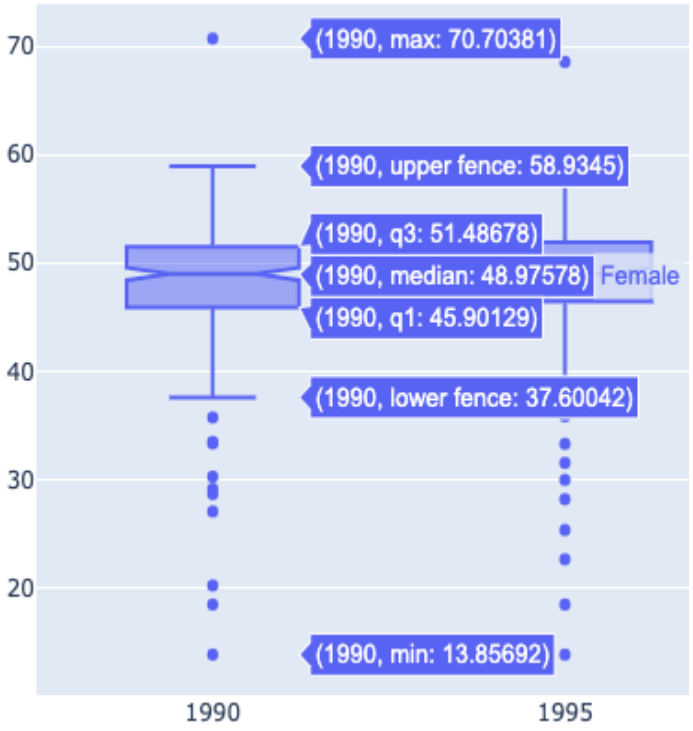
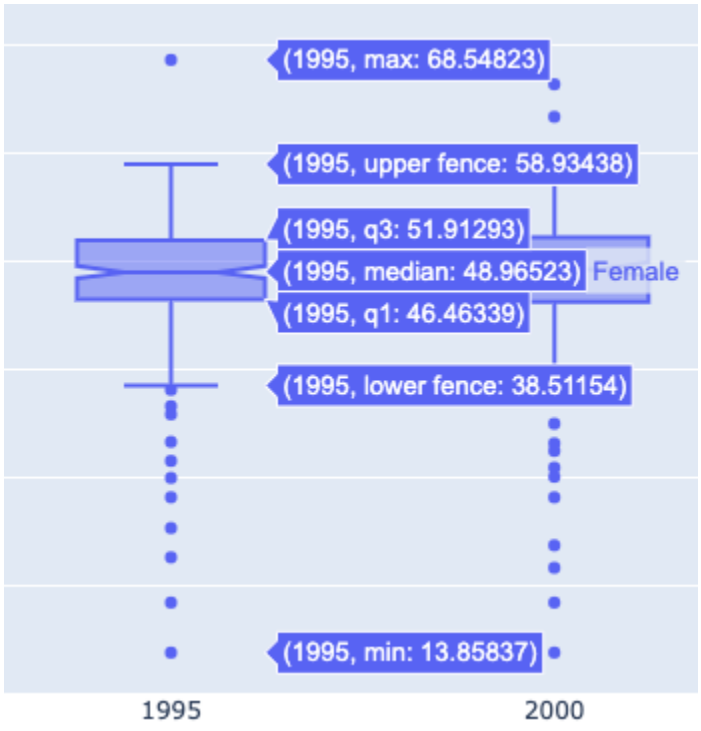
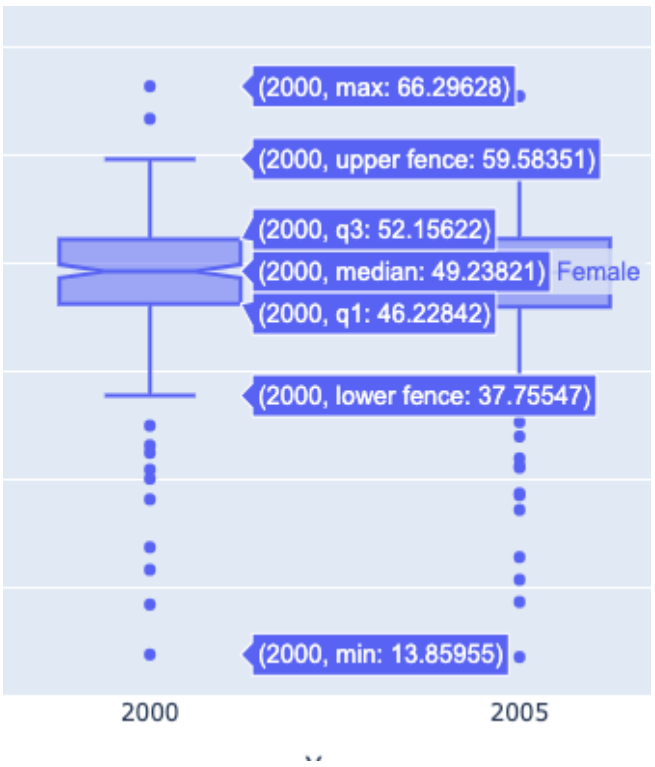


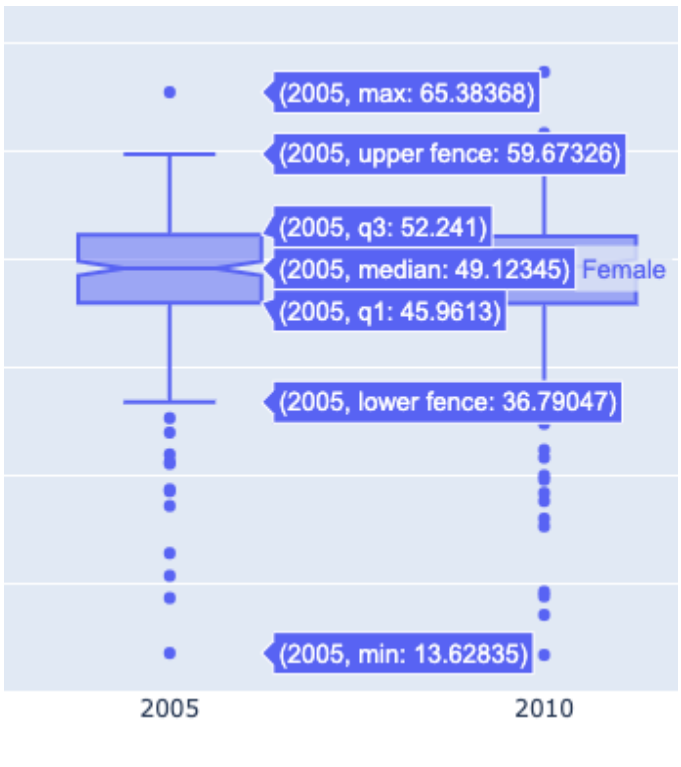
Figure 4: Boxplot graph for female migrants as a percentage of international migrants. The X-axis is the year from 1990 to 2015 and the y-axis shows female migrants as a % of

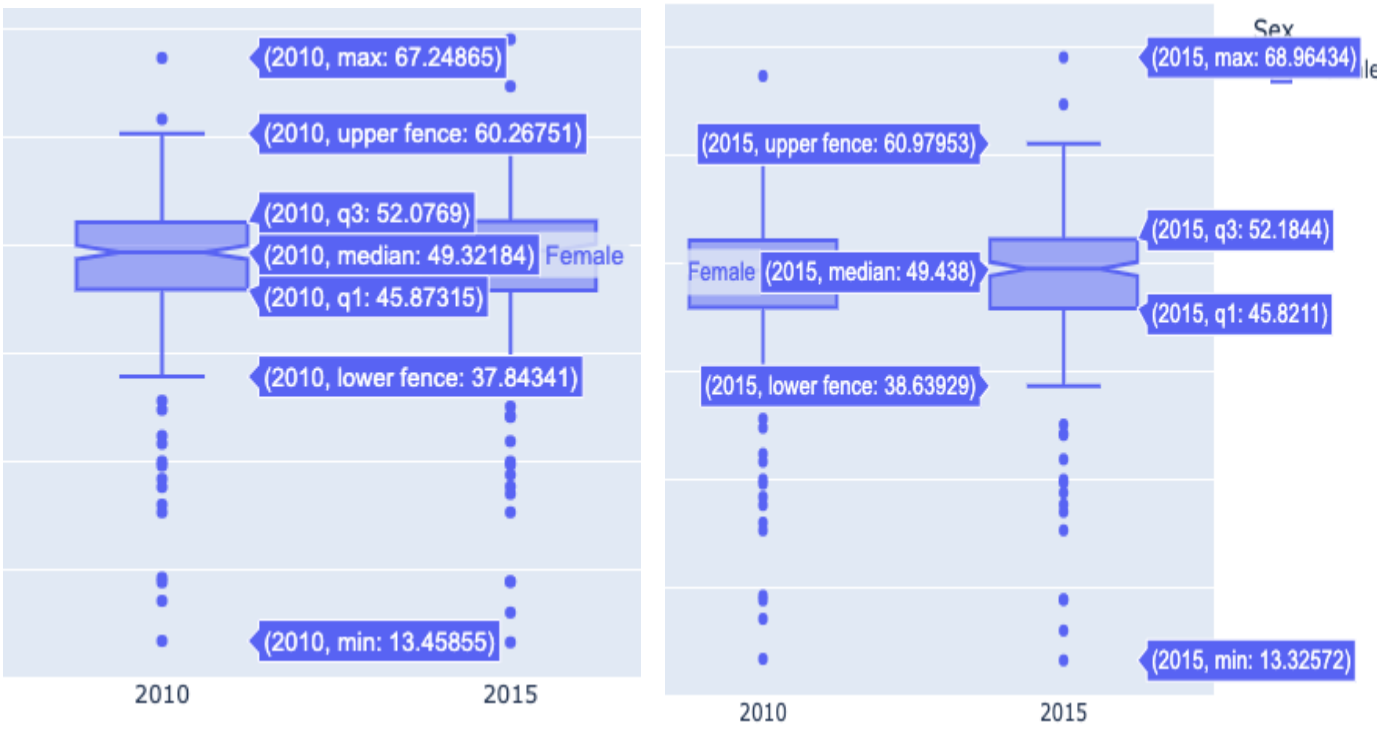
international migrant stocks. We can also see the outlier countries in the graph are Nepal being the max outlier and Bangladesh being the min outlier.











Screenshots from figure 4: Boxplot graph data for individual years. Data points include minimum, lower fence, quartile 1, median, quartile 3, upper fence, and maximum for each year from 1990 to 2015.

5) The fifth subset of data is on the annual rate of change of migrant stock from 1990 to 2015 including both female and male migrants. This data was visualized using a violin plot graph where it is showing female and male data side by side for a particular time frame. The x-axis shows the time frame in years and the y-axis shows the annual rate of change of the migrant stock. This was done based on Tufte's principle of comparison, and causality, and included multivariate data. In this one graph, users can quickly compare the female and male annual rates of change side by side and form a correlation between time and rate of change of migrants. Each violin plot also can provide data points such as minimum, lower fence, quartile 1, median, quartile 3, upper fence, and maximum. Please refer to the chart for specific data points. Overall, it was very clear that during 1990 - 1995 there was the largest annual change in migrant stock. For all time frames, the overall median annual rate of change is between 1%-2%.

Anja Zhang
INF1340 - Final Assignment Write-up

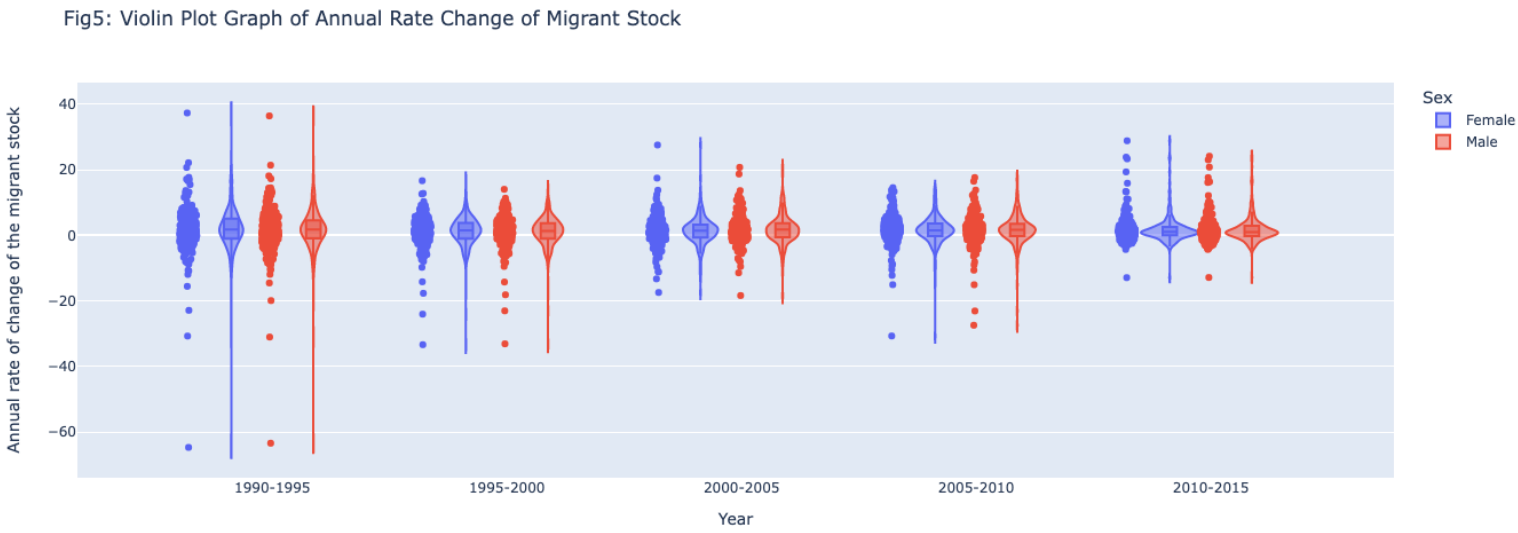


Figure 5: violin plot graph of the annual rate of change for migrant stock. The timeframe in years is in the x-axis and the annual rate of change of migrant stocks in percentage is in the y-axis. Female and male data are separated using blue and red colors.

6) There are three subsets of data which was produced from table 6. They are Estimate refugee stock at mid-year, refugee as a percentage of the international migrant stock, and Annual rate of change of the refugee stock. For each dataset, there are the corresponding year and countries with each data point. To accurately visualize the dataset, it is important to create separate visualization to properly interpret each data. Using Tufte's principles, show comparison, show causality, and show multivariate three charts were created. They are an animated scatter graph for estimated refugee stock at mid-year for figure 6.1, a line graph for refugees as a percentage of the international migrant stock for figure 6.2, and a scatter graph for the annual rate of change of the refugee stock. For all graphs, the x-axis is the time in years, and the y-axis is the corresponding metrics for each dataset. The main intention for each graph is to easily compare the between each country at the same time, users can easily pick up trends. More result analysis will be provided below each graph.

Anja Zhang
INF1340 - Final Assignment Write-up

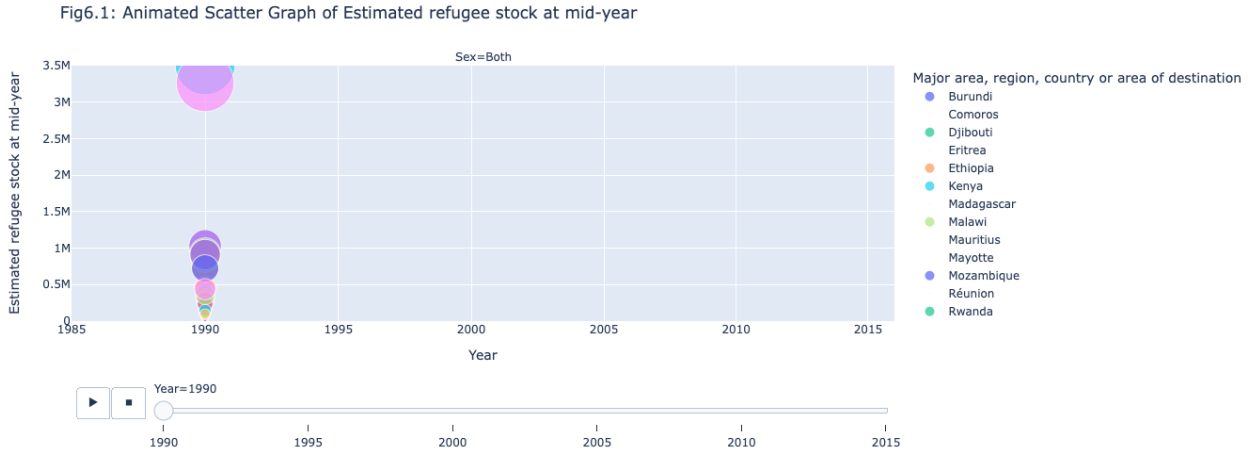


Figure 6.1: Scatter graph of estimated refugee stock at mid-year. The X-axis is the time in years and the y-axis displays refugee stocks. Each country is represented by a color and the size of the circle corresponds to the number of refugees. Users can quickly perform an analysis based on the visualization by looking at the country with the most refugee for each year. 1990 is Iran, 1995 is Iran, 2000 in Pakistan, 2005, 2010, and 2015 arinJoran. With this information, users can quickly speculate there is political unrest at the time and use this data as the starting point for further analysis. Its animated feature can also show the fluctuation of refugees in each country through time, and it's very clear that there are fluctuations in refugees numbers.

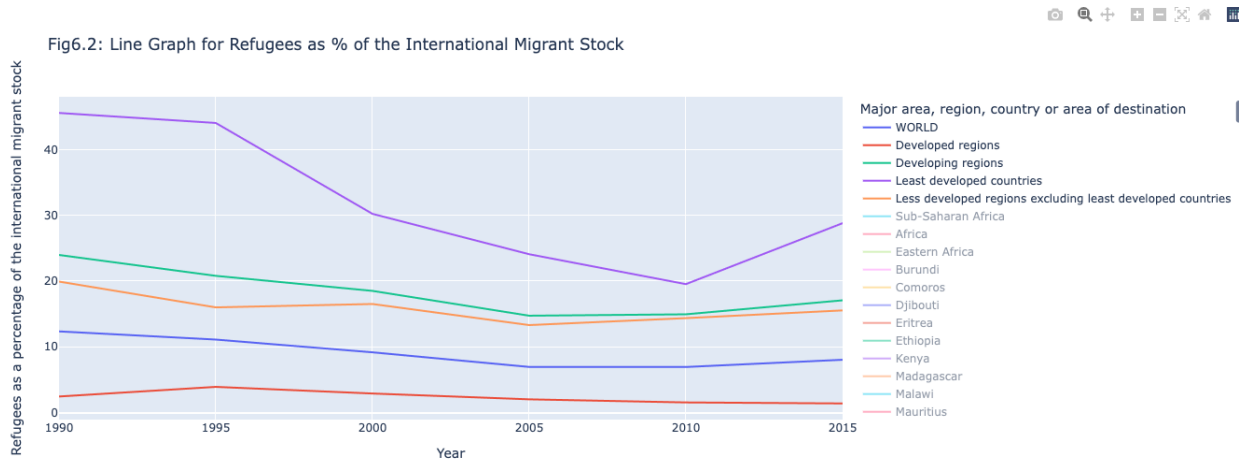


Figure 6.2: Line graph for refugees as a percentage of the international migrant with data shown per country. The x-axis is the time of year and the y-axis is the refugees as a percentage of the international migrant stock. Each color and line represent data for a particular country or region. The above graph is built on Tufte's principle of comparing, as this can showcase an overall global trend through time but more importantly, this graph makes it very easy for users to compare certain countries or regions side by side. As the figure shows, the refugees are drastically different between developed and

developing regions. Developed regions have a refugee % below 5% for all years from 1990 to 2015. In contrast, developing regions' percentages are from 15% to 25%. This means that there is a strong relationship between the economy of the country and its refugee numbers.

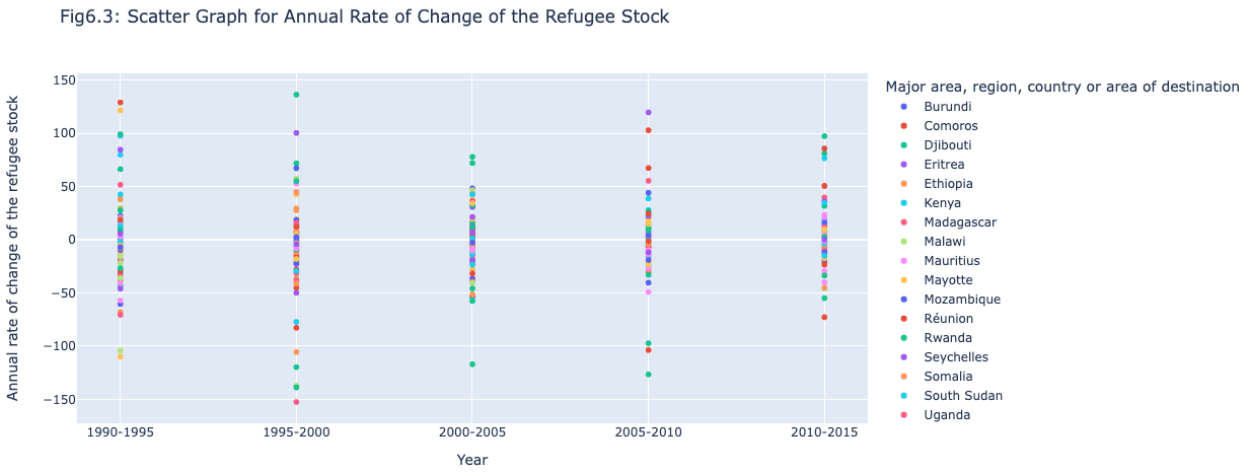


Figure 6.3: scatter graph for the annual rate of change of refugee stock. The x-axis is the time in the year and the y-axis is the annual rate of change of the refugee stock. Each country is represented with a different color. Using the graph above, users can easily interpret which country has the highest and the lowest annual rate of change of refugees. In this case, they are Afghanistan from 1990-1995, Indonesia from 1995-2000, Bahrain from 2000-2005, Venezuela from 2005-2010, and Niger from 2010-2015. This data can be a starting point to look into what are the reason for this high spike in refugee numbers for a particular time frame.

Discussion

This project developed used various visualization tools to better display data from the “Trend in Internation Migrant Stock: the 2015 Revision” dataset from United Nations. Tufte’s principles were central to this process. They include Showing comparisons, Showing causality, showing multivariate data, Use appropriate encodings, Use integrated captions, and Show Context. I believe this comprehensive list of principles is a good guideline to follow when making decisions on which visualization to utilize. The goal of visualization graphs is to convey information to the stakeholders, Tufte’s principle helps with this process because it allows the data scientist to better think through what information should be made apparent.

One reflection which I discover from creating this project is that the quality of data is very closely related to the quality of visualization that can be created. As a data scientist, it is important to have the skillset for both data cleaning and data visualization. Furthermore, to make the data

visualization more relevant, it would be beneficial for data scientists to learn about the business need of the data to make better data-related decisions.